Stanford Project: PROTEIN STRUCTURE MODELING (CRYSALIS)

Principal Investigator: Edward A. Feigenbaum, Ph.D.
Department of Computer Science
Stanford University

Contact: Allan TERRY@SUMEX-AIM
(415) 497-1740

The CRYSALIS system is an application of artificial intelligence methodology to the task domain of protein crystallography. The focus is the structure determination problem: the derivation of an atomic model of the protein from an indistinct image of the electron density. The crystallographer interprets these data in light of the known chemical composition of the protein, general principles of protein chemistry, and his own experience. The goal of the CRYSALIS Project is to integrate these diverse sources of knowledge and data into a program that matches the crystallographer's level of performance in electron density map interpretation. A successful solution to this problem must deal with issues such as representation and management of a large knowledge base, opportunistic reasoning, and appropriate description of the emerging hypothesis, while keeping human engineering considerations in sight. Automation of this task would shorten the time for protein determination by several weeks to several months and would fill a major gap in the construction of a fully-automated system for protein crystallography.

SOFTWARE AVAILABLE ON SUMEX

CRYSTALLOGRAPHIC DATA REDUCTION PROGRAMS (in FORTRAN):
- A density map skeletonizer (SKEL37) based on an improved version of Greer's algorithm.
- A package for locating the critical points in a map.
- A general map-manipulation utility (INSPCT) that can find peaks, display regions, and compute various statistics.

TWO LISP SYSTEMS (with the caveat that both are under active development):
- A system (SEGLABELING) which heuristically parses the segmented map into labels similar to those a crystallographer would use.
- The inference system (CRYSALIS).

REFERENCES

Engelmore, R.S. and Nii, H.P.: A knowledge-based system for the interpretation of protein x-ray crystallographic data. Heuristic Programming Project Report HPP-77-2, Computer Science Dept., Stanford Univ., January, 1977.

Engelmore, R. and Terry, A.: Structure and function of the CRYSALIS system. Proc. Sixth IJCAI, Tokyo, August, 1979.

Nii, H.P. and Feigenbaum, E.A.: Rule-based understanding of signals. Heuristic Programming Project Report HPP-77-7, Computer Science Dept., Stanford Univ., April, 1977.

Stanford Project: RX--DERIVING KNOWLEDGE FROM TIME-ORIENTED CLINICAL DATABASES

Principal Investigators: Robert L. Blum, M.D.
Departments of Medicine
and Computer Science
Stanford University
Stanford, California 94305
(415) 497-3088 (BLUM@SUMEX-AIM)

Gio C.M. Wiederhold, Ph.D.
Departments of Computer Science
and Electrical Engineering
Stanford University
Stanford, California 94305
(415) 497-0635 (WIEDERHOLD@SUMEX-AIM)

The objective of clinical database (DB) systems is to derive medical knowledge from the stored patient observations. However, the process of reliably deriving causal relationships has proven to be quite difficult because of the complexity of disease states and time relationships, strong sources of bias, and problems of missing and outlying data.

The goal of the RX Project is to explore the usefulness of knowledge-based computational techniques in solving this problem of accurate knowledge inference from non-randomized, non-protocol patient records. Central to RX is a knowledge base (KB) of medicine and statistics, organized as a taxonomic tree consisting of frames with attached data and procedures. The KB is used to retrieve time-intervals of interest from the DB and to assist with the statistical analysis. Derived knowledge is incorporated automatically into the KB. The American Rheumatism Association DB containing 7,000 patient records is used.

SOFTWARE AVAILABLE ON SUMEX

RX--(excluding the knowledge base and clinical database) consists of approximately 200 INTERLISP functions. The following groups of functions may be of interest apart from the RX environment:

SPSS Interface Package: Functions which create SPSS source decks and read SPSS listings from within INTERLISP.
Statistical Tests in INTERLISP: Translations of the Piezer-Pratt approximations for the T,F, and Chi-square tests into LISP.
Time-Oriented Data Base and Graphics Package: Autonomous package for maintaining a time-oriented database and displaying labelled time-intervals.

REFERENCES

Blum, R.L. and Wiederhold, G.: Inferring knowledge from clinical data banks utilizing techniques from artificial intelligence. Proc. Second Annual Symposium Computer Applications in Medical Care, IEEE, Washington, D.C., November, 1978, pp. 303-307.

Blum, R.L.: Automating the study of clinical hypotheses on a time-oriented database: The RX project. Submitted to MEDINFO80, Third World Conference on Medical Informatics, Tokyo, 1980.

Weyl, S., Fries, J., Wiederhold, G. and Germano, F.: A modular self-describing clinical databank system. Comp. and Biomed. Res. 8(3):279-293, June, 1975.

Wiederhold, G., Fries, J.F.: Structured organization of clinical data bases. AFIPS Conference Proc. 44:479-485, 1975.

Appendix B

Resource Operations and Usage Statistics

The following data give an overview of various aspects of SUMEX-AIM resource usage. There are five sub-sections containing data respectively for:

1) Overall resource loading data

2) Relative system loading by community

3) Individual project and community usage

4) Diurnal loading data

5) Network usage data

6) System reliability data

1. Overall resource loading data

The following plots display several different aspects of system loading over the life of the project. These include total CPU time delivered per month, the peak number of jobs logged in, and the peak load average. The monthly "peak" value of a given variable is the average of the daily peak values for that variable during the month. Thus, these "peak" values are representative of average monthly loading maxima and do not reflect the largest excursions seen on individual days, which are much higher.

These data show well the continued growth of SUMEX use and the self-limiting saturation effect of system load average, especially after installation of our overload controls early in 1978. Since late 1976, when the dual processor capacity became fully used, the peak daily load average has remained between about 5.5 and 6. This is a measure of the user capacity of our current hardware configuration and the mix of AI programs.

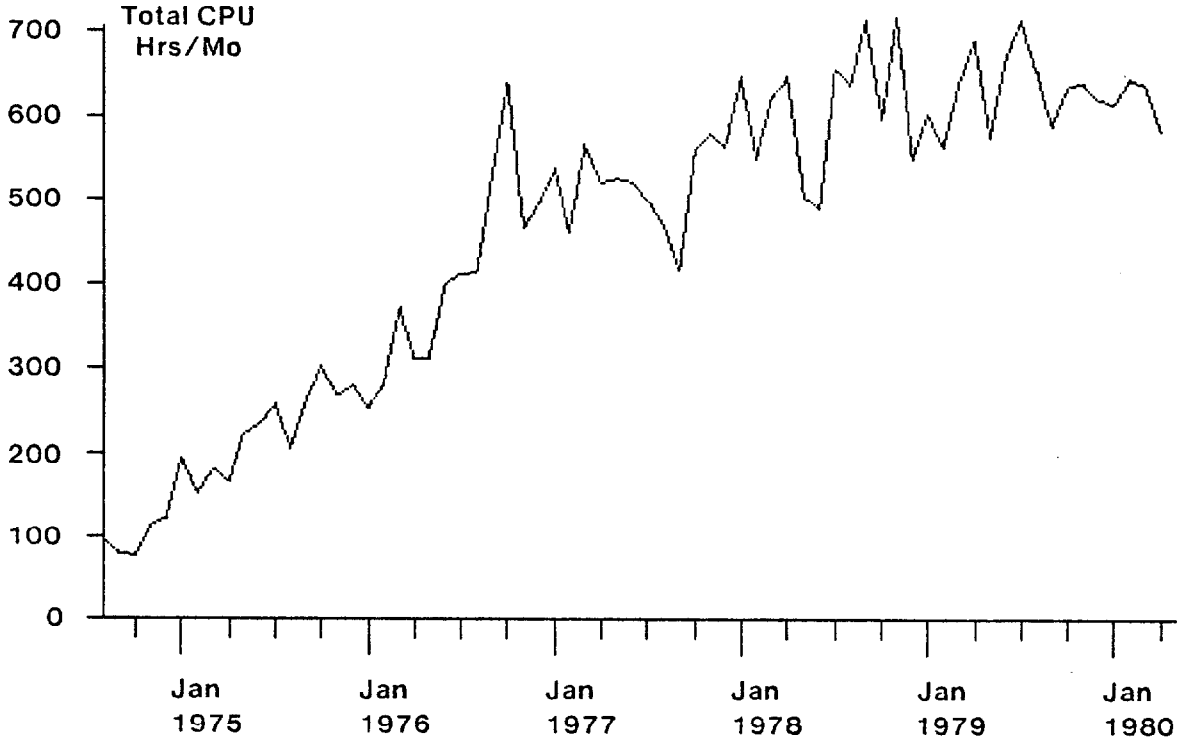

Figure 7. Total CPU Time Consumed by Month

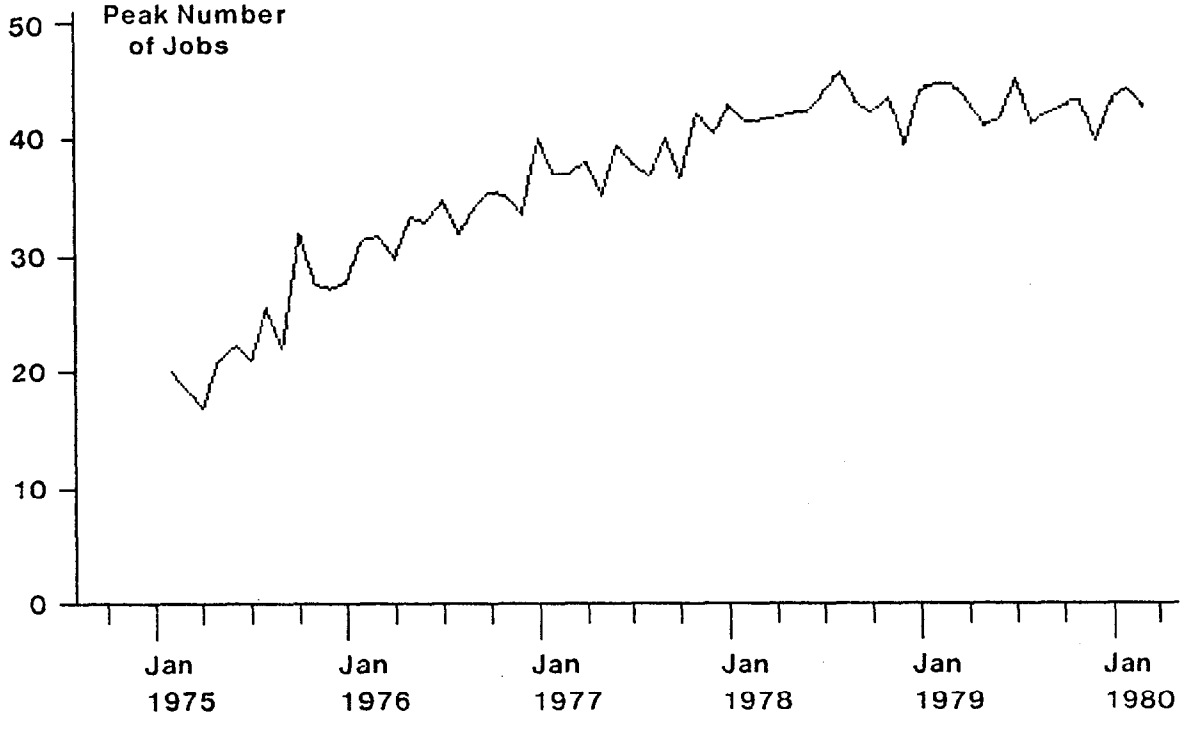

Figure 8. Peak Number of Jobs by Month

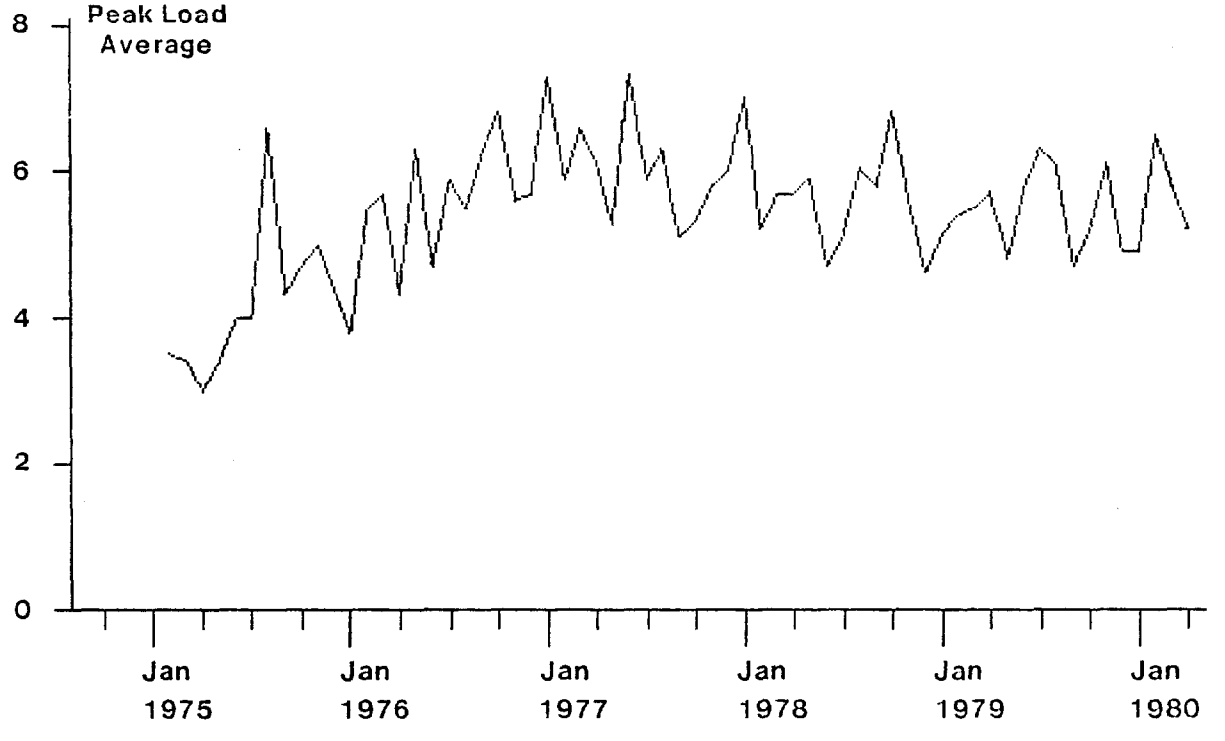

Figure 9. Peak Load Average by Month

2. Relative System Loading by Community

The SUMEX resource is divided, for administrative purposes, into 3 major communities: user projects based at the Stanford Medical School, user projects based outside of Stanford (national AIM projects), and common system development efforts. As defined in the resource management plan approved by BRP at the start of the project, the available system CPU capacity and file space resources are divided between these communities as follows:

Stanford	40%
AIM	40%
Staff	20%

The "available" resources to be divided up in this way are those remaining after various monitor and community-wide functions are accounted for. These include such things as job scheduling, overhead, network service, file space for subsystems, documentation, etc.

The monthly usage of CPU and file space resources for each of these three communities relative to their respective aliquots is shown in the plots in Figure 10 and Figure 11. Terminal connect time is shown in Figure 12. It is clear that the Stanford projects have held an edge in system usage despite our efforts at resource allocation and the substantial voluntary efforts by the Stanford community to utilize non-prime hours. This reflects the maturity of the Stanford group of projects relative to those getting started on the national side and has correspondingly accounted for much of the progress in AI program development to date.

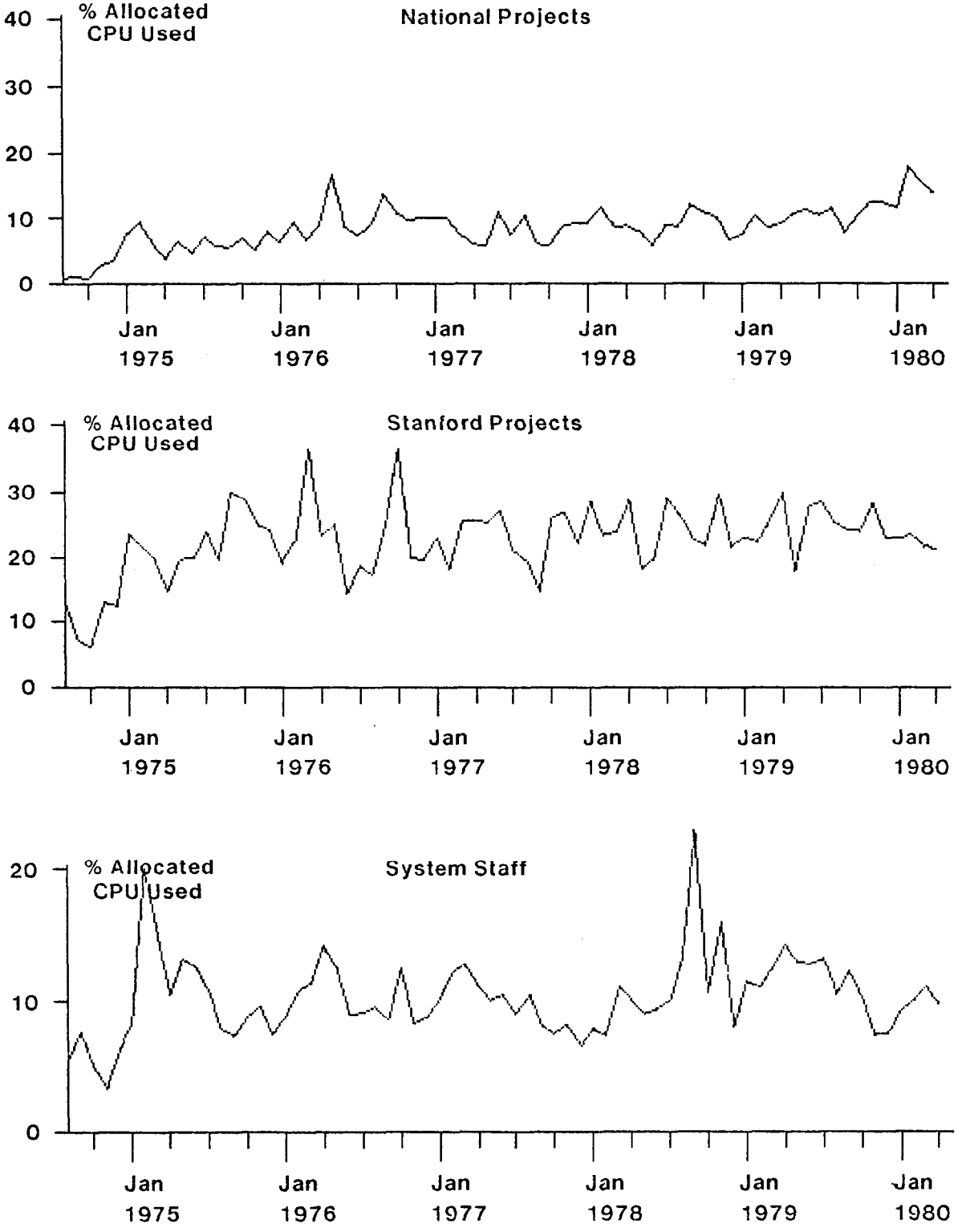

Figure 10. Monthly CPU Usage by Community

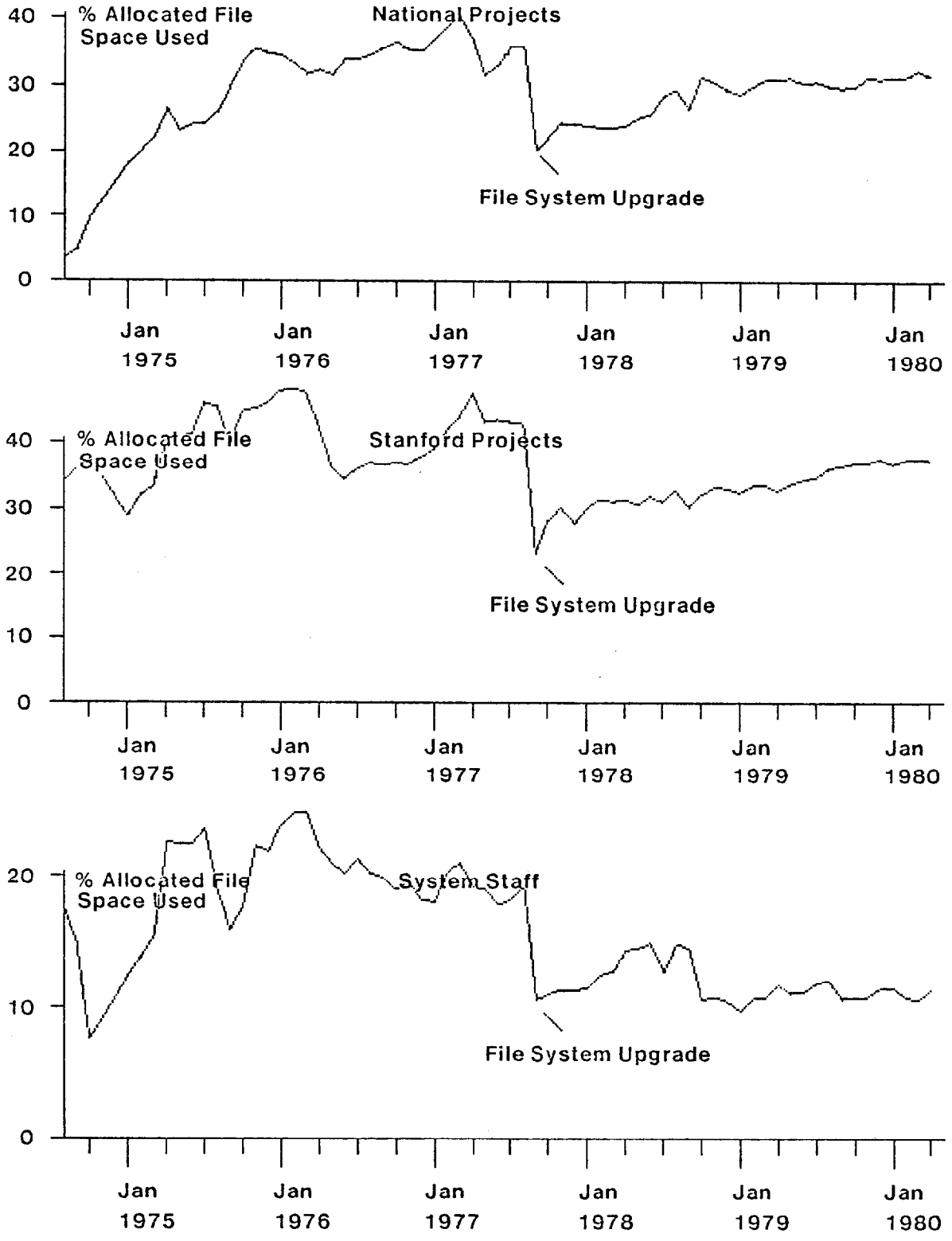

Figure 11. Monthly File Space Usage by Community

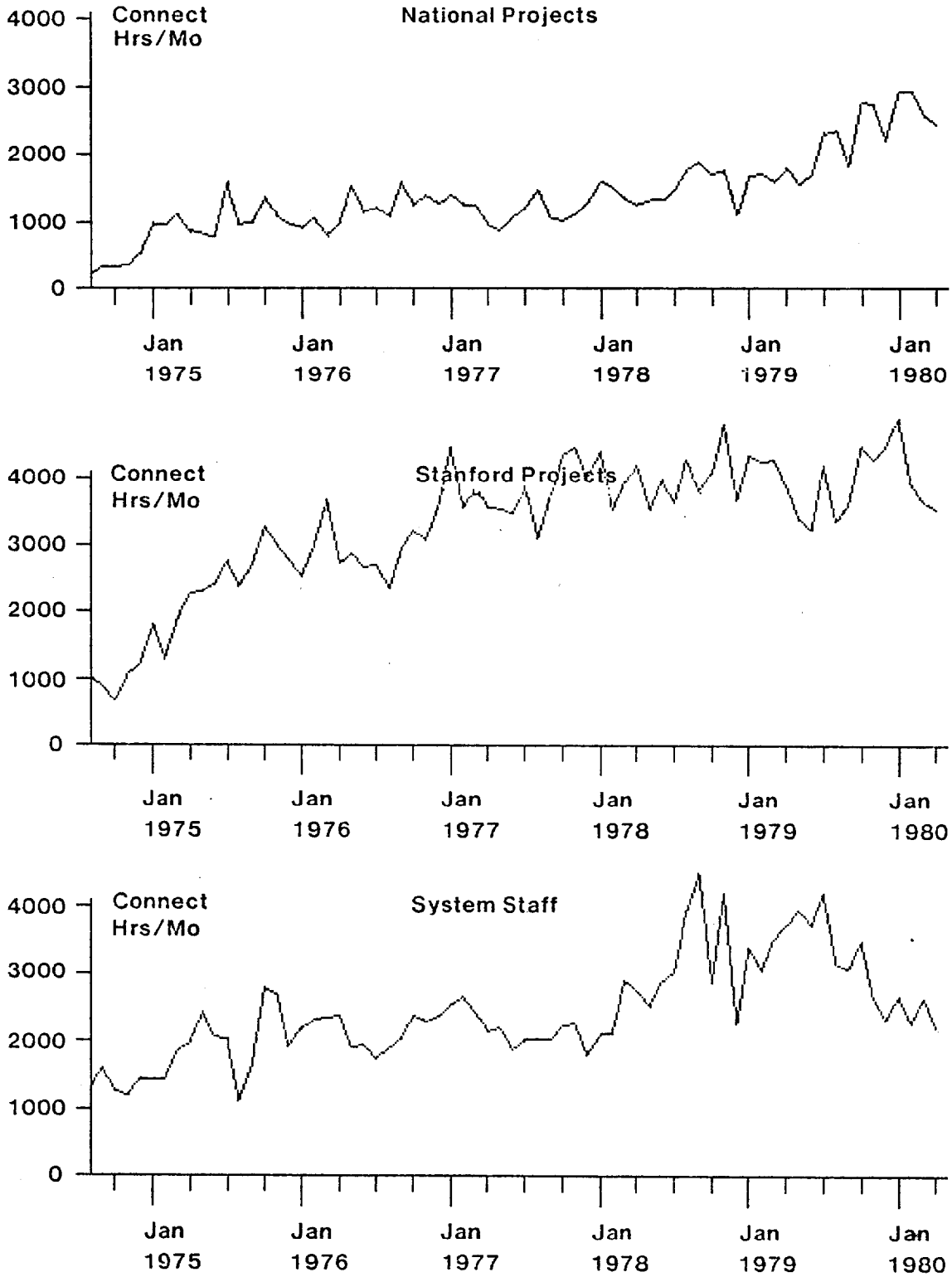

Figure 12. Monthly Terminal Connect Time by Community

3. Individual Project and Community Usage

The table following shows cumulative resource usage by project during the past grant year. The entries include a summary of the operational funding sources (outside of SUMEX-supplied computing resources) for currently active projects, total CPU consumption by project (Hours), total terminal connect time by project (Hours), and average file space in use by project (Pages, 1 page = 512 computer words). These data were accumulated for each project for the months between May 1979 and April 1980.

Several of the projects newly admitted to the National AIM community use the Rutgers-AIM resource as their home base. These projects are listed in the tables to fully document the scope of the AIM community and are noted with the flag "[Rutgers-AIM]".

Again the well developed use of the SUMEX resource by the Stanford community can be seen. It should be noted that the Stanford projects have voluntarily shifted a substantial part of their development work to non-prime time hours which is not explicitly shown in these cumulative data. It should also be noted that a significant part of the DENDRAL, MYCIN, AGE, AI Handbook, and MOLGEN efforts, here charged to the Stanford aliquot, support development efforts dedicated to national community access to these systems. The actual demonstration and use of these programs by extramural users is charged to the national community in the "AIM USERS" category, however.

Resource Use by Individual Project - 5/79 through 4/80

National AIM Community	CPU (Hours)	Connect (Hours)	File Space (Pages)
1) ACT Project "Acquisition of Cognitive Procedures" John Anderson, Ph.D. Carnegie-Mellon Univ. ONR N00014-77-C-0242 9/78-9/80 $175,000	106.50	1197.90	2634
2) SECS Project "Simulation & Evaluation of Chemical Synthesis" W. Todd Wipke, Ph.D. U. California, Santa Cruz NIH RR-01059-03S1 (3.7 yrs. 7/77-2/81) 7/80-2/81 $36,949 NIH/NCI NO1-CP-75816 (2 yrs. 1/79-12/80) 1/80-12/80 $74,394	538.31	9943.77	8389
3) Mod Human Cogn Project "Hierarchical Models of Human Cognition" Peter Polson, Ph.D. Walter Kintsch, Ph.D. University of Colorado NIE-G-78-0172 (3 yrs. 9/78-8/81) 9/79-8/80 $46,537 NIMH MH-15872-9-13 (5 yrs. 6/76-5/81) 6/79-5/80 $32,880 ONR N00014-78-C-0433 6/78-5/80 $68,315 6/80-5/81 $60,000 ONR N00014-78-C-0165 (2 yrs. 1/78-12/80) 1/80-12/80 $85,000	119.61	2696.51	712
4) Higher Mental Functions "Intelligent Speech Prosthesis" Kenneth Colby, M.D. UCLA NSF MCS-78-09900 6/78-11/80 $135,260 NSF PFR-17358 10/79-3/81 $318,368	20.65	637.47	2810

5)	INTERNIST Project "DIALOG: Computer Model of Diagnostic Logic" Jack Myers, M.D. Harry Pople, Ph.D. University of Pittsburgh NIH RR-01101-03 (3 yrs. 7/77-6/80) 7/79-6/80 $200,414	215.76	3756.99	7755
6)	PUFF/VM Project "Biomedical Knowledge Engineering in Clinical Medicine" John Osborn, M.D. Inst. Medical Sciences, San Francisco Edward Feigenbaum, Ph.D. Stanford University NIH GM-24669 9/78-8/81 $164,000 (*) Supplement pending	125.39	4669.66	3196
7)	SCP Project "Simulation of Cognitive Processes" James Greeno, Ph.D. Alan Lesgold, Ph.D. University of Pittsburgh NIE-G-80-0114 (3 yrs. 12/79-11/82) 12/79-11/80 $217,000 ONR/ARPA N00014-79-C-0215 (1.8 yrs. 1/79-9/81) 10/79-9/80 $420,000 NSF/NIE 12/78-5/81 $161,238 ONR N00014-78-C-0022 (3 yrs. 10/77-9/80) 10/79-9/80 $92,293	21.07	648.56	764
8)	*** [Rutgers-AIM] *** Rutgers Project "Computers in Biomedicine" Saul Amarel, D.Sc. NIH RR-00643 (3 yrs. 12/77-11/80) 12/79-11/80 $451,383	23.56	513.01	9204

9)	*** [Rutgers-AIM] ***			
	Decision Models in			
	Clinical Diagnosis	.00	.00	0
	Robert Greenes, M.D.			
	Harvard University			
	NLM LM-03401			
	(5 yrs. 7/79-6/84)			
	7/79-6/80 $235,582			
10)	*** [Rutgers-AIM] ***			
	Heuristic Decisions in			
	Metabolic Modeling	.00	.00	0
	David Garfinkel, Ph.D.			
	Univ. Pennsylvania			
	HL-15622			
	(3 yrs. 12/77-11/80)			
	12/79-11/80 $111,051			
	GM-16501-11A1			
	(2 yrs. 4/80-3/82)			
	4/80-3/81 $60,598			
	Proposals pending			
11)	AIM Pilot Projects			
	Coagulation Expert	9.67	207.67	480
	Commun. Enhancement	1.99	85.45	361
	KRL Demonstrations	.36	11.53	523
	MISL Project	2.40	115.23	1132
	Psychopharm. Advisor & Statistical Advisor	6.11	183.49	818
	Refinement of Med. Know.	.00	.00	0
	Struct. for Med. Diag. [Rutgers-AIM]	.45	10.85	0
	AIM Pilot Totals	20.98	614.22	3320
12)	AIM Administration	16.67	661.64	4668
13)	AIM Users on Stanford Projects			
	AGE	3.26	57.94	35
	DENDRAL	140.61	1930.07	1592
	MOLGEN	20.53	293.11	106
	MYCIN	11.75	391.32	167
	Guest (all projects)	34.28	362.49	209
	Other	1.78	27.31	230
	AIM User Totals	212.21	3062.24	2342
	Community Totals	1420.71	28401.97	45794

Stanford Community	CPU (Hours)	Connect (Hours)	File Space (Pages)
1) AGE Project (Core) "Generalization of AI Tools" Edward Feigenbaum, Ph.D. ARPA MDA-903-80-C-0107 (**) (partial support)	341.46	3103.55	3277
2) AI Handbook Project (Core) Edward Feigenbaum, Ph.D. ARPA MDA-903-80-C-0107 (**) (partial support)	69.34	2149.53	2611
3) DENDRAL Project "Resource Related Research Computers and Chemistry" Carl Djerassi, Ph.D. NIH RR-00612-11 (3 yrs. 5/80-4/83) 5/80-4/81 $221,255	957.35	10625.34	15112
4) MOLGEN Project "Experiment Planning System for Molecular Genetics" Edward Feigenbaum, Ph.D. Laurence Kedes, M.D. NSF MCS-78-02777 12/79-11/80 $153,959 (*)	409.20	9229.85	7242
5) MYCIN Project "Computer-based Consult. in Clin. Therapeutics" Bruce Buchanan, Ph.D. Edward Shortliffe, M.D., Ph.D. NLM LM-03395 (5 yrs. 7/79-6/84) 7/79-6/80 $99,484 NSF MCS-79-03753 7/79-12/80 $146,152 ONR/ARPA N00014-79-C-0302 3/79-3/82 $396,325 (*) NLM LM-00048 (5 yrs. 7/79-6/84) 7/79-6/80 $39,285 Kaiser Fdn. 7/79-12/80 $20,000 (*)	632.87	11594.76	12809

6)	Protein Struct Modeling "Heuristic Comp. Applied to Prot. Crystallog." Edward Feigenbaum, Ph.D. NSF MCS-79-23666 12/79-11/81 $35,318	85.84	1631.07	4443
7)	RX Project Robert Blum, M.D. Gio Wiederhold, Ph.D. Pharm. Mfr. Assn. Fdn. 7/78-6/80 $32,500 NLM New Invest. 7/79-6/82 $90,000 NCHSR 4/79-3/81 $35,000 Proposal pending	21.03	817.30	1256
8)	Stanford Pilot Projects			
	Genetics Applic.	55.98	938.73	377
	Hydroid	30.34	1202.01	1037
	Ultrasonic Imaging	18.67	331.44	309
	Miscellaneous	.00	.00	8
	Stanford Pilot Totals	104.99	2472.18	1732
9)	Stanford and HPP Assoc.	211.83	6710.84	6827
	Community Totals	2833.91	48334.42	55309

SUMEX Staff		CPU (Hours)	Connect (Hours)	File Space (Pages)
1)	Staff	900.39	27809.06	9731
2)	MAINSAIL Development	272.64	5526.13	3493
3)	Staff Associates, misc.	45.42	1850.25	3249
	Community Totals	1218.45	35185.44	16473

System Operations	CPU (Hours)	Connect (Hours)	File Space (Pages)
1) Operations	2088.14	84421.51	75174
	=======	=========	======
Resource Totals	7561.21	196343.34	192750

* Award includes indirect costs. All other awards are reported as total direct costs only.

** Supported by a larger ARPA contract MDA-903-80-C-0107 awarded to the Stanford Computer Science Department:

	Current Year (10/79-9/80)	Total Award (10/79-9/82)
Heuristic Programming Project	$ 496,256	$1,613,588
VLSI/CAD Network	248,918	685,374
Total award	$ 745,174	$2,298,962

4. System Diurnal Loading Variations

The following figures give a picture of the recent variations in diurnal SUMEX system load, taken during April 1980. The plots include:

Figure 13 - Total number of jobs logged in to the system

Figure 14 - System load average (average number of simultaneously runnable jobs)

Figure 15 - Percent of total CPU time used by logged in jobs (maximum is 200% for dual processor capacity)

The abscissa for these plots is broken into 20 minute intervals throughout the day. The ordinate for each interval is the average of all the daily measurements for that interval over the weekdays during April 1980. A daily measurement for a given 20 minute interval is in turn an average of the appropriate statistic sampled every 10 seconds. Since these plots display overall average data, they give representative illustration of the general characteristics of diurnal loading. There are, of course, substantial fluctuations in the quantities measured from day to day as well and for some, also on time scales shorter than the intervals displayed in the figures. For example in Figure 14, the number of runnable jobs shows a fairly smooth curve peaking at 5.2 jobs. On both a scale of minutes and from day to day, however, the number of runnable jobs will vary from only a few to 12 or more. These fluctuations are not shown in these average plots but also play an important role in the responsiveness of the system.

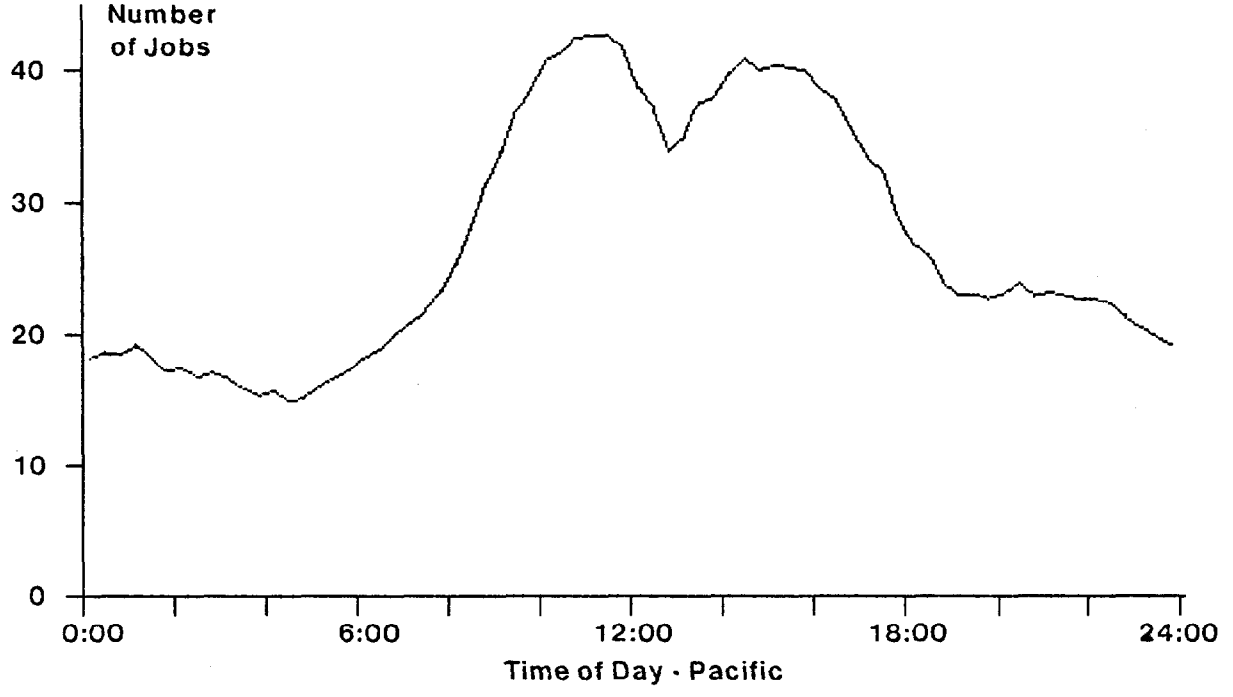

Figure 13. Average Diurnal Loading (4/80): Number of Jobs

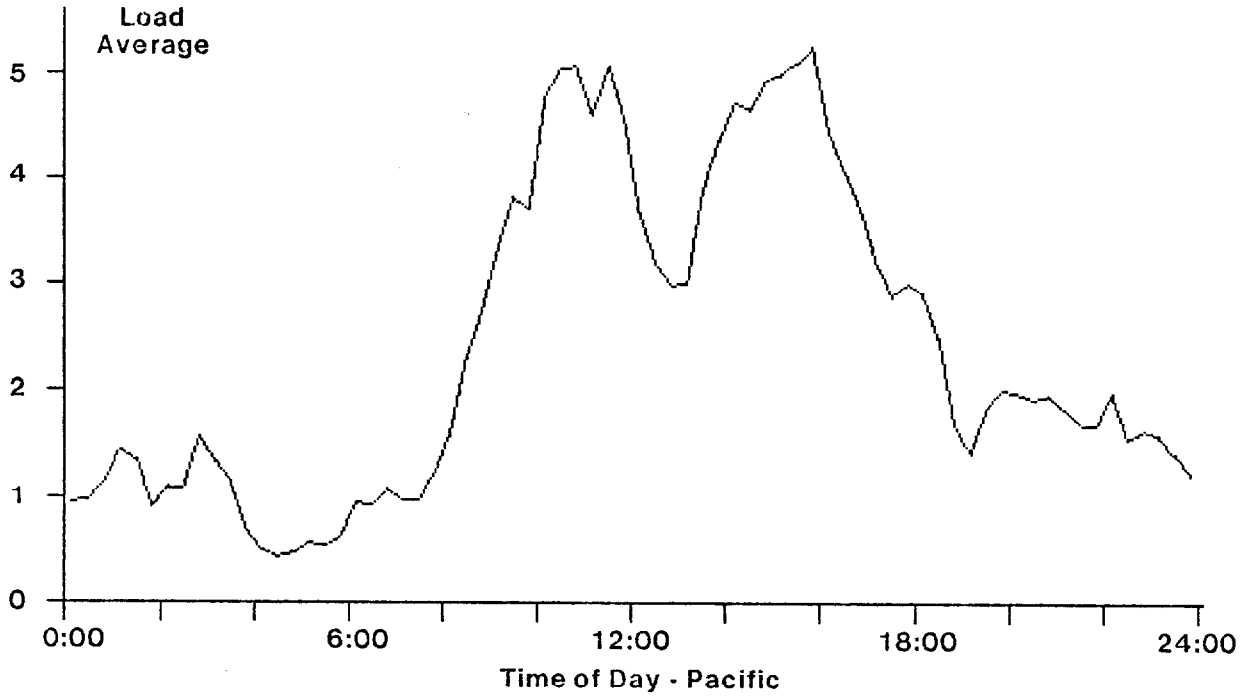

Figure 14. Average Diurnal Loading (4/80): Load Average

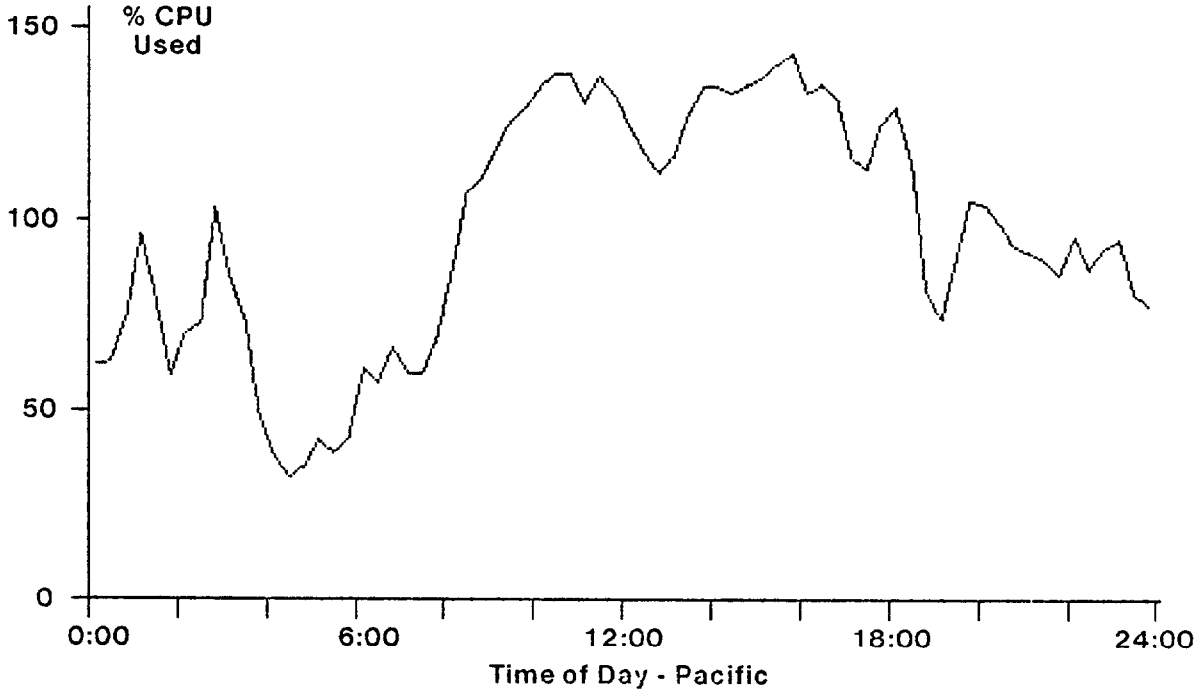

Figure 15. Average Diurnal Loading (4/80): Percent Time Used

5. Network Usage Statistics

The plots in Figure 16 and Figure 17 show the monthly network terminal connect time for TYMNET and ARPANET. This forms the major billing component for SUMEX-AIM TYMNET usage. The terminal connect time does not reflect the time spent in file transfers and mail forwarding.

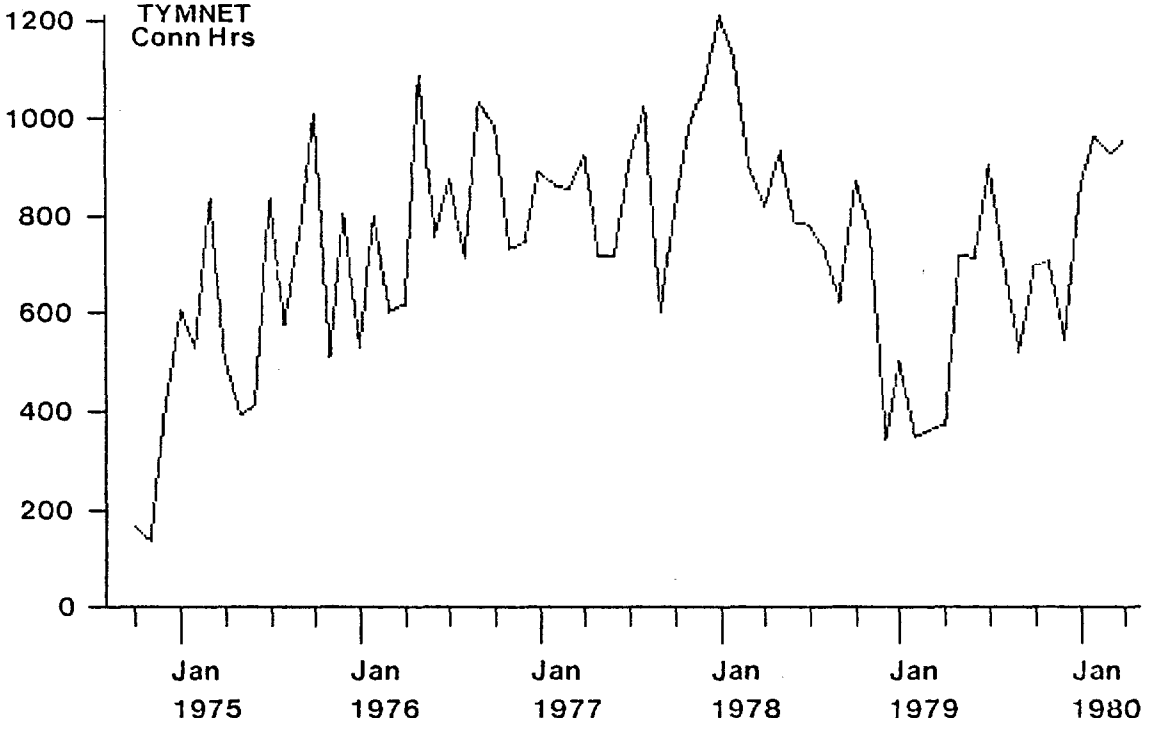

Figure 16. TYMNET Terminal Connect Time

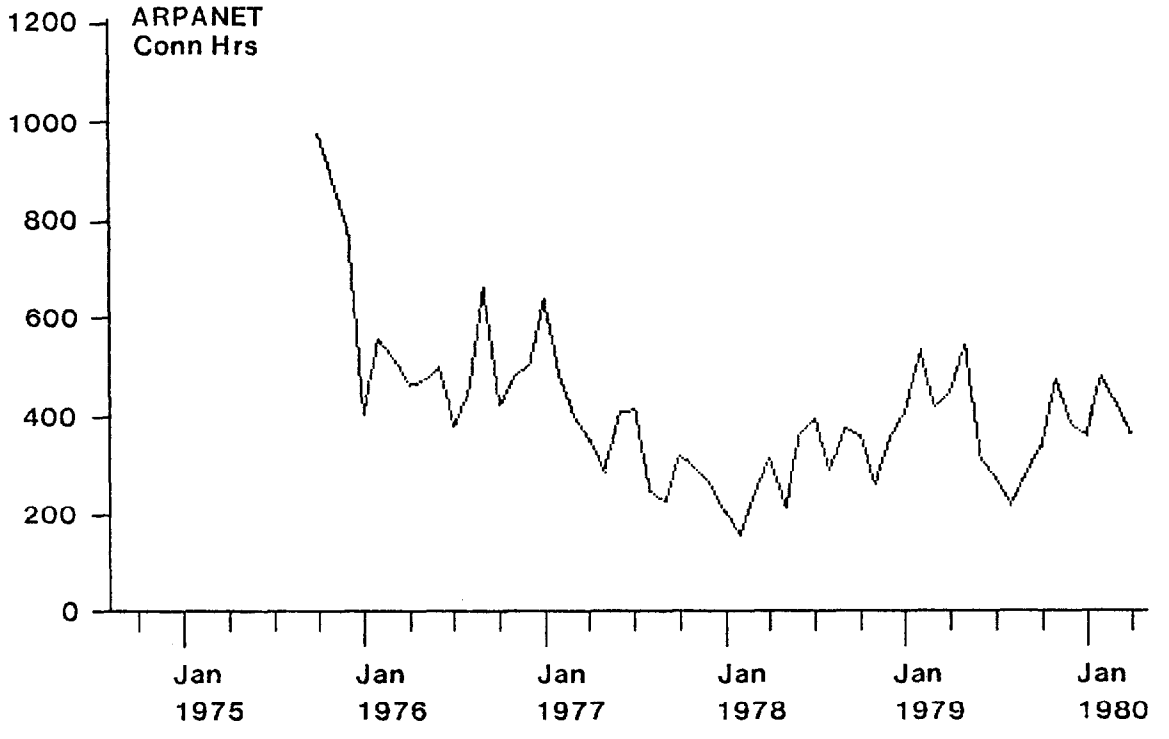

Figure 17. ARPANET Terminal Connect Time

6. System Reliability

System reliability has been very good on average with several periods of particular hardware or software problems. The table below shows monthly system reloads and downtime for the past year. It should be noted that the number of system reloads is greater than the actual number of system crashes since two or more reloads may have to be done within minutes of each other after a crash to repair file damage or to diagnose the cause of failure.

	1979								1980			
	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR
RELOADS												
Hardware	12	4	1	15	1	0	13	6	4	2	9	8
Software	0	3	5	1	7	3	2	1	4	0	6	4
Environmental	1	1	0	0	0	1	0	1	0	1	0	2
Unknown Cause	0	2	3	1	3	0	0	2	0	0	0	1
Totals	13	10	9	17	11	4	15	10	8	3	15	15
DOWNTIME (Hrs)												
Unscheduled	38	18	15	12	18	4	33	28	8	4	14	38
Scheduled	19	28	20	35	38	29	19	15	28	27	41	23
Totals (Hrs)	57	46	35	47	56	33	52	43	36	31	55	61

TABLE 1. System Reliability by Month

Appendix C

Local Network Integration

The introduction of satellite machines into the SUMEX facility raises important issues about how best to integrate such systems with the existing machines. We seek to minimize disruptions to the operational resource with the addition of new machines, the duplication of peripheral equipment, and the interdependence among machines that would increase failure modes. We also require high-speed intermachine file transfer capabilities and terminal access arrangements allowing a user to connect flexibly to any machine of choice in the resource.

The initial design of the SUMEX system was that of a "star" topology centered on the KI-10 processors. In this configuration, all peripheral equipment and terminal ports were connected directly to the KI-10 buses. With the addition of satellite machines, a unique focus no longer exists and some pieces of equipment need to be able to "connect" to more than one host. For example, a user coming into SUMEX over TYMNET will want to be able to make a selection of which machine he connects to. Another TYMNET user may want to make another choice of machine and so the TYMNET interface needs to be able to connect to any of the hosts. This could be accomplished by creating separate interfaces for each of the hosts to the TYMNET, each with a different address. Besides being expensive to duplicate such interfaces, it would be inconvenient for a user to reconnect his terminal from one host to another. He would have to break his existing connection and go through another connect/login process to get to another machine. Since we want to facilitate user movement between various machines in the SUMEX resource, this process needs to be as simple as possible - in fact a user may have jobs running simultaneously on more than one machine at a time.

Similarly, we need to be able to quickly transfer files between any two machines in the resource, connect common peripheral devices (e.g. printer or plotter) to any machine desiring to use them, and allow any host to access other remote resources such as Stanford campus printers or terminal clusters. If we were to establish direct connections pairwise between machines and devices, the number of such connections would go up quadratically with the number of devices.

A more effective solution lies in the implementation of a local network in which all devices (host CPU's, peripheral devices, network gateways, etc.) are tied to a shared communications medium and can thereby establish logical connections as needed between any pair of nodes. Such network systems have been under development for a number of years, taking on various topological configurations and control structures depending on bandwidth requirements and interdevice distances. A very attractive design for a highly localized system configuration from the viewpoint of simplicity, reliability, and bandwidth is the Ethernet which has been under development for several years at Xerox Palo Alto Research Center [10]. The

Ethernet utilizes a fully distributed control structure in that each device connected to the net can independently decide to send a message to any other device on the net depending on the functions it is actively performing. Of course, decisions about which devices need to communicate with each other at a given time and what the precise message content is are determined by higher level system activities and requests, for example to implement a file transfer, mail forwarding, teletype connection, printer output, etc. Current Ethernets operate at 3 Mbits/sec and realize over 90% effective capacity utilization under heavy load [11]. Protocols exist is handle "collisions" between two devices trying to gain control of the network at the same time and to interconnect Ethernet with other networks.

The Stanford Computer Science Department is one of three recipients of grants from Xerox that includes Ethernet connection, terminal, and graphics printer equipment. Since the Computer Science Department systems are integrally connected with one of the major user groups on SUMEX (the Heuristic Programming Project) and since the Ethernet design is ideal for the the integration of new satellite machines with the existing SUMEX facility, we have chosen it as the model for our planned facility changes.

A diagram of the on-going Ethernet implementation for SUMEX is shown in Figure 3. Plans include developing interfaces for each host machine, the TYMNET, the local teletype scanner, other peripheral devices, and a gateway to other local networks (e.g., the Computer Science Department machine and planned terminal clusters). We already have the KI-10's connected through an I/O bus interface and are almost ready to debug the 2020 interface. These both use the Xerox interface board designed for PDP-11's. We are also working on a more efficient connection for the KI-10's through a direct memory access device and on connections for the other resources.